A METHOD FOR CONTENT MINING OF SEMI-STRUCTURED DOCUMENTS

ABSTRACT OF THE INVENTION

Embodiments of the present invention are directed to a method for content mining of semi-structured documents. In one embodiment, a semi-structured document is first converted from a document-type specific format such as HTML or PDF, to a document-type independent format such as XML. The document formatting, which contains basic level information about the document's structure, is then analyzed by a series of modules to develop a higher level understanding of the document's structure. These modules append information to the document describing the features which collectively comprise the higher level document structure. The appended information facilitates finding specified information within the document when content mining is performed.